自然言語処理 78－18
（1990．7．20）

# 英語ニュースの機械翻訳

浦谷則好　　相沢輝昭
NHK放送技術研究所

ＮＨＫでは衛星放送で英語ニュースのテロップ作成に英日機械翻訳システムを試用している．ニュース文は扱う分野が広い，複雑な固有名詞表現や数量表現が多い等の特徴がある．実際に機械翻訳で使用された英語ニュース文を分析したところ，１文当り１１．０語で，現在時制の平叙文が多いこと，無生物主語の文が多いこと等が判明した．また，機械翻訳での解析成功率は６４．５％であった．衛星放送のニュース文を対象としていることによって生ずる点を中心に我々のシステムの問題点についても報告する．現在，英語ニュースデータベースを構築し，これを用いて辞書・文法の見直しを進め翻訳精度の向上を図っている．

## English－Japanese Machine Translation for News

Noriyoshi Uratani,　　Teruaki Aizawa

NHK Science and Research laboratories

1-10-11 Kinuta, Setagaka-ku, Tokyo, 157 Japan

NHK has been trying a machine translation system from English to Japanese for producing subtitles for news. News sentences have complex proper nouns and numeral expressions. We analyzed the English news sentences which were translated by our system. Most of them were declarative sentences with the present tense whose average length was 11.0 words. The success rate of translation was 64.5 percent. The problems of our system are described around the use in TV news. A database for AP news is being constructed. We are going to make use of it to improve the grammar and the dictionary of our MT system.

## 1．はじめに

機械翻訳システムは多くの電算機メーカー，ソフトウェアハウス，各種研究機関などにおいて研究開発が進められており，既に販売されているものも少なくはない．しかし，その多くは製品マニュアル，技術文献を対象にしており，翻訳精度も十分なものではない．

ＮＨＫでは，ニュースという広い分野を対象として，放送における機械翻訳技術の実用化を目指して，昭和６３年度から英日機械翻訳の研究を行っている．現在，すでに衛星放送でテロップ（字幕）作成に試験的に使用しつつ[1]，問題点の抽出，その改善を図っている．

以下，衛星放送における機械翻訳の試用状況や問題点を具体的に例示した後，今後の改良の進め方を述べる．また，翻訳精度向上のために構築している英語ニュースデータベースにもふれる．

## 2．衛星放送における機械翻訳システムの試用状況

ＮＨＫにおける翻訳の需要は多い．例えば，衛星第１放送の「ワールドニュース」では，英・仏・独・伊・露・韓・中の各言語で入って来る外国語ニュースに迅速に日本語テロップを付けて放送しなければならない．この翻訳作業は時差の関係で深夜作業となることも多く，運用上の大きなネックとなっている．

機械翻訳システムは，この部分の将来における省力化を目的に導入され，昨年８月から英語ニュースの日本語テロップの作成に試用されている．ニュースの聴取から英文の要約までは人手で行う．テロップは通常１５字程度であり，最大でも３０字なので，この段階で相当大幅な要約が必要となる．この際，併せて英文の単純化などの前編集もなされる．この要約英文の日本語への翻訳を機械翻訳システムが担当する．翻訳された結果には後編集がなされ，日本語テロップが作成され，画面に合わせて送出される．（写真１）

放送用英日機械翻訳システムの現在の諸元は表１のようになっている．翻訳の基本ソフトウェアとしては既存のもの[2]を導入した．

## 3．英語ニュースの機械翻訳の分析結果

衛星第１放送「ワールドニュース」のために実際に機械翻訳に供された英語ニュース文（４５日分）の分析を行った．結果を以下に示す．


写真１　　実際の放送画面

表１　　システム諸元

| | |
|---|---|
| 翻訳方式 | トランスファ方式 |
| 辞書の語数 | 基本語４万語 |
| | 専門語５．５万語 |
| 構文解析規則 | 約３０００ |
| ハードウェア | ＳＵＮ３／２６０ |
| 使用言語 | Ｃ |
| 翻訳速度 | 約２万語／時 |

### 3．１　「ワールドニュース」文の特徴

対象ニュース文の各種統計は以下のようになった．

| | | |
|---|---|---|
| 総文数 | １，３９３ | |
| 総語数 | １５，４２２ | |
| | （ピリオド，コンマ除く） | |
| ・語数／文 | １１．０ | |
| ・動詞／文 | １．９ | |
| 異なり語数 | ３，３８９ | （数値を除く） |
| ・固有名詞 | ２９１ | |
| ・人名 | １２０ | |

文型は平叙文が１，０８８（７８％）で最も多く，次はタイトル等の無動詞文（１５７：１１％）であった．時制に関しては，主文が現在のものが９０２であり，補文でも現在時制が２７１と過去，未来に比して圧倒的に多く使用されていた．これは，画面と文が連

動しているテロップ文の特徴であろう．また，受動態の使用は１３４であった．本動詞以外の動詞の主な使われ方は，不定詞２５７，進行形１１２，完了形１０９であった．

その他では，無生物を主語とする文が５０１と多いことが目立った．これは翻訳対象としたニュースが政治経済が少なく，市民社会のトピック的なものが多かったためだろうと考えられる．また，従属節は２２１，ＡＮＤ並列は２２０，関係節は１８１出現し，これも比較的多いと言えよう．

### ３．２　システムの翻訳レベルの現状

「ワールドニュース」の1,393文に対する機械翻訳結果は次の通りである．

| | |
|---|---|
| 全　文　数 | １，３９３ |
| 解析に成功 | ８９８ (64.5%) |
| 第１候補 | ６９８ (77.7%) |
| 第２候補 | １２１ |
| 第３候補 | ７０ |
| 解析に失敗 | ４９５ |
| スペルミス | １３９ (28.1%) |
| 非　文 | １２ |

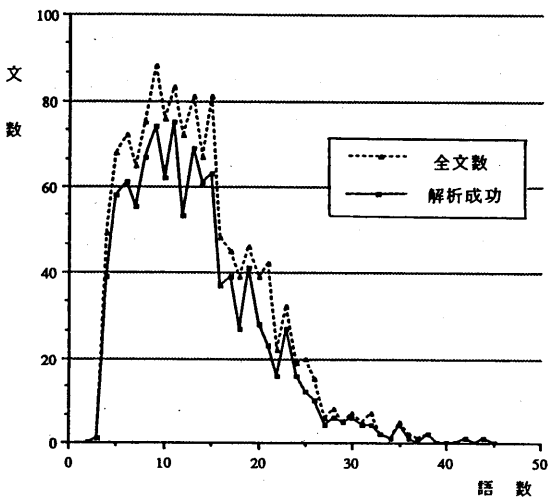


図１　語数と解析成功数との関係

文の語数と解析成功文の数との関係を図１に示す．

この結果から多くの候補文から最適の文を選択する機構（ウェイト機構）はおおよそうまく機能していると言えよう．解析に失敗したものの約３０％は原文のスペルや非文が原因であり，それを除けば解析成功率は７２．３％ (898/1242) となる．

代表的な翻訳例を次に挙げる．

This is a new program exploring solutions to the most pressing environmental and social problems of our time.

㊊これは我々の時間の最も緊急の環境の，そして社会的問題の解を探求している新しいプログラムである．

㊐並列句のスコープも正しく認定している．

Nearly 600 British soldiers and 300 police have been killed.

㊊ほとんど６００イギリスの軍人と３００警察は死んだ．

The Cheyenne Indians claim their ancestors fled to west before Columbus landed because prophecy warned them to flee the coming of a white destruction.

㊊予言がそれらに白い破壊の接近を避けなさいと警告したので，Columbus が上陸する前に，それらの先祖が西へ逃げたと Cheyenne インディアンは主張する．

He plays easy listenig music as they munch on celery stocks.

㊊それらがセロリ株でむしゃむしゃ食べるように，彼はイージーリスニング音楽を演奏する．

㊐easy listenig music が辞書に登録されているのでうまく訳されている．これが辞書に無いと「容易な聞いている音楽」と訳されて何のことか分からなくなる．

次に，我々の機械翻訳システムの問題点を，いくつかの翻訳結果をもとに，以下に例示する．ただし，品詞認定，訳語選択，ｔｏ不定詞の訳し分け等のどの翻

訳システムにも共通する問題は除外した．

★衛星放送ニュースを対象としたことから起こった問題点

◆画面を見ないと文意が分からないものがある．

They slither like a slity toves

㊐彼らは slity toves を slither 好む．

㊎この文はニュースキャスターがslitherからの連想で「鏡の国のアリス」のJabberwockyの詩の一節を思わず口ずさんだもの．この文だけだとミススペルとしか思えない．文脈処理を考えても対処不能と思われる．なお，slitherは未登録語であり，品詞認定を誤っている．

◆対象とする分野が広く，かつ個別には専門的．

（しかも学問的な専門分野と異なりトピック的）

Free-lance writer, Chuck Young, is trying to do purple haze just like Jimi Hendrix did it.

㊐フリーの作家（Chuck Young）はJimi Hendrixのようなもやだけがそれにした紫色を行うことを…

㊎purple haze が曲名（「紫のけむり」）であると知らないと正しく訳せない．

◆間をカンマで表わすと，通常の構文解析が失敗．

"The new Tears For Fears album is called, "the seeds of love."

㊐新しいTears For Fears アルバムが呼ばれる‖，‖「愛の種」

㊎カンマが無ければもちろん正訳が得られる．

★前編集が不十分なために生じた問題点

◆ミススペルに弱い．

They began to disappear as land values shut up.

㊐彼らは土地価値が締まるように見えなくなり始めた．

㊎shut up は shot up（急騰する）の誤り．

スペルチェック機能があってもユーザは使ってくれると限らない！

◆口語的な表現が多い．

"This organized crime element, Russian emigrants, they kill anybody."

㊐「この組織された罪元素‖（ロシアの移民）彼らは誰も殺す．」‖

◆省略に弱い．

"Why not buy a motorized wheel chair?"

㊐「Whyが自動車化された車輪いすを買わない‖？」

◆文の分割に対する制約が厳しい．

While Irish fighters and Irish civilians, Catholic or Protestant, who have died, were 2000.

㊐アイルランドのが死んだ武人かアイルランドの一般国民かカトリックかプロテスタントは2000であった．

㊎従属節だけの文となっている．主文と一緒にすると正しい訳が得られた．

★その他

◆複合語，イディオムの登録が不十分

| | | |
|---|---|---|
| light house | 誤訳：光家 | 正訳：灯台 |
| air line | 誤訳：空気ライン | 正訳：航空会社 |
| be scheduled to | 誤訳：表に組み入れられている | |
| | 正訳：することになっている | |

◆複雑な数量表現への対処が不十分

A family of four can see a show for under 10 …

㊐４の１人の家族が，ショーを見ることができ…

この他にも係り先の認定，並列句のスコープ解釈，同格挿入句の解釈，分節構造の認定等の誤る例が見られる．しかし，これらは頻度も多くなく，また他の機械翻訳システムに比して劣っているとも考えられないので割愛した．さらに，助詞等の不適切な使用によって適切な日本語訳とならない例も多いが，今回は構文解析レベルを中心に報告したいのでこれも省く．

## 4．翻訳精度改善のための英語ニュースデータベース

我々のシステムは放送用の機械翻訳システムとして，まだ十分とは言えない．ユーザーインターフェイスの改良としては，引き続き翻訳操作の改善や新語登録機能の充実などを図っていくつもりである．また機械翻訳システムと放送システムとのリンクもこれからの課

題である．テロップ送出装置とのオンライン接続など様々な形で機械翻訳を活用すべく検討を進めている．

3章で見てきたように我々のシステムの翻訳性能もまだ十分とは言えない．そこで，ニュース文では特徴的な固有名詞表現や数量表現の解析強化を先行して進める[3]とともに，以下のような改善を進めている．

## 4．1　辞書の充実

機械翻訳システムは2種類の辞書を使い分けるようになっている．1つは，各語について詳細な記述を持つ基本語辞書であり，もう1つは基本的に英日の対訳のみからなる専門語辞書である．

基本語辞書は当初2万語であったが，学習研究社の「学研英和電子辞書」を追加して4万語に拡張している．1つの見出し語には複数の訳語が対応しており，訳語の選択には文型パターンや意味マーカーなどを用いている．しかし，現状では特殊な語を除けば意味マーカーは付与されていない．そこで，約80項目からなる意味マーカー体系を構成し，各語への付与を進めている．[4]

一方，専門語の方は，試用開始当初はゼロの状態であった．そこで，次の3種類の冊子型辞書を用いて充実を図り，現在5．5万語になっている．[5]

三省堂『ニューズ英語辞典』

小学館『最新英語情報辞典』

集英社『情報・知識ｉｍｉｄａｓ』

## 4．2　英語ニュースデータベースの構築

機械翻訳システムの翻訳精度の向上のためには，対象とする実際のデータに即して，辞書や文法をチューニングしていく必要がある．このため、我々はＡＰ電を中心にニュースの実データの収集を行っている．ＡＰ電は昨年2月以降，ＡＰ通信社から直接電話線を介して入電しており，パソコンを用いて自動収集している．1日当り約350記事（約700ＫＢ）が得られている．これを，効率良く利用するためにはデータベース化は不可欠である．

この英語ニュースデータベースの利用目的としては一応以下のものを想定している．

1）ニュース文そのものの検索

2）ニュース文からの専門語・慣用句（固有名詞表現，数値表現を含む）の抽出

3）ニュース文からの表現パターンの抽出

4）辞書・文法の評価

データベース上のデータ表現としては，データの管理，保守，各種の処理を考えれば，関係データベースやオブジェクト表現が有利であろうが，データの分析（特にミクロな分析）の手法が固まっていない現在ではむしろできるだけ生データに近い形にして後の処理に制約を加えたくないと考えた．そこで，単に記事毎にヘッダー，本文，トレイラを分離し，図2のように各記事のヘッダー，トレイラの部分と本文部分をまとめて格納することとした．図2で，“￥￥”はヘッダー部とトレイラ部を分けるデリミタであり，“＃＃”は記事本文やヘッダー・トレイラ部の各々を分けるデリミタである．さらに，データの先頭には記事毎のヘッダー・トレイラ部へのポインターと本文部へのポインターを付加した．ＡＰ電は日毎に，ワールドニュースは月毎にこのデータ形式に変換することにした．このような簡単な形式にしたので，原データに対してわずか1％の容量増で済ますことができた．自動収集されたＡＰ電は変換コマンドによって自動的にＥＷＳ上のデータベースに取り込んでいるが，回線ノイズ等によってデータの欠落や，不正データの混入が起こることがあり，この場合はデータチェック等の支援コマンドを用いて人手で修正してデータベース化している．現在（90.6）ＡＰ電のデータは246ＭＢ格納されている．処理の基本となる文字列の検索は当所で開発したＦＡＳＴ法[6]を組み込んで高速に実行している．

本格的なデータベースの利用に先立って，ＡＰ電データの予備調査を行った．ＡＰ電は1日に約350記事（約2ＫＢ／記事）送られてくるが，この中には要約記事，改作記事等の重複記事が多く含まれている．このままで各種統計をとっても有用なデータにはなりえない．さらに，相場記事，スポーツ記事は一般ニュースとは別個に扱いたい．そこで，47個のキーワードを設定し，これをヘッダー中に含む記事をＦＡＳＴ法で探索し，不要と思われる記事を統計の対象から除外した．

1989年4月～90年3月（データの欠落がある

ため実質約２９０日分）の約１６００万語について調査してみた．異なり語数は約１７万３千語であった．出現語数に対する異なり語数の変化を図３に，出現頻度に対する累積出現率の変化を図４に示す．上位３０位までの高頻出単語は表２に示す．

先に述べた衛星放送の分析と，ＡＰ電の調査から以下の特徴が判明している．

＜衛星放送／ＡＰ電に共通＞

・使われている用語が広範囲に及び，しかも分類が難しい．

・人名，機関名，国名，地名などの固有名詞が頻出する．人名は肩書を含んだ複雑な表現が多い．

・複雑な数量表現が頻出する．

```
address      内    容
(16進)
00000     1c20   172f9              記事#1へのポインター
00010     1ca1   1849e              記事#2へのポインター
            :      :
01c10     17225  d719f              最後の記事へのポインター
01c20  W1001-----u IBX   W0002  25-04 00756   }
       233 87 10 lds ldp snp txm BOM            }
       ^^AP News Digest,0789<                   } 記事#1のヘッダー
       ^EDITORS:<                               }
       ¥¥                                       }
       ^END<                                    } 記事#1のトレイラ
       AP-TK-25-04-90 0004GMT<                  }
       ##
01ca1  W1002-----u SBX   W0003  25-04 00384   }
       233 87                                   }
       ^^Nicaragua-Contras,0429<                } 記事#2のヘッダー
       ^LaserPhotos available<                  }
            :                                   }
       ##
172f9     MOSCOW - The KGB has reinforced its   }
       units on the Lithuanian border and       } 記事#1の本文
       increased patrols offshore, an off...    }
                  :                             }
          By Mary McVean.                       }
       ##
1849e   Joaquina Garcia is a peasant woman      }
       who hasn't seen four sons since they     } 記事#2の本文
       left to join the Contra rebels nine      }
                  :                             }
d7b33  ##
```

図２　英語ニュースデータベースのデータ形式

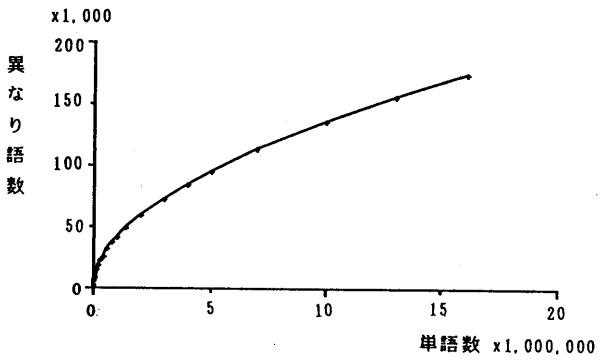


図３　出現語数に対する異なり語数の変化

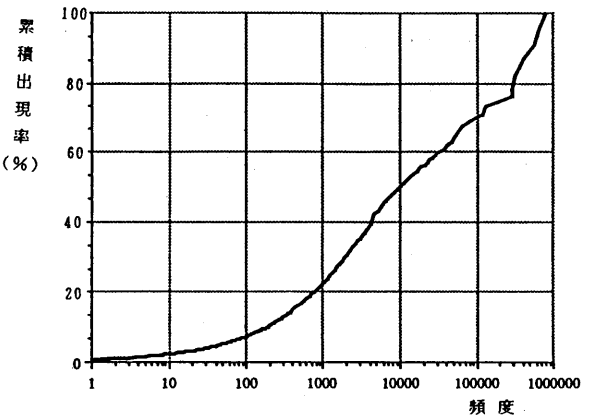


図４　出現頻度に対する語の累積出現率の変化

＜ＡＰ電のニュース＞

・１文が比較的長い．（２３．７語／文）

・言動・思考に関する動詞が多用される．特に say は be動詞に次いで使われ have動詞より頻度が多い．call, talk, report, ask, think, want, feel の使用も多い．

・時制は「過去」の文が多い．

表２　ＡＰ電の高頻度単語（上位３０位）
（１６００万語中，　単位千語）

| | | | |
|---|---|---|---|
| the | 982 | on | 109 |
| ,(カンマ) | 749 | he | 89 |
| .(ﾋﾟﾘｵﾄﾞ) | 592 | by | 87 |
| of | 424 | with | 85 |
| to | 373 | will | 78 |
| a | 373 | it | 71 |
| in | 347 | from | 69 |
| be動詞 | 344 | at | 68 |
| and | 328 | as | 65 |
| 数字 | 304 | but | 52 |
| say | 201 | not | 48 |
| have動詞 | 159 | his | 47 |
| 's | 137 | government | 44 |
| for | 132 | who | 42 |
| that | 123 | U.S. | 36 |

＜衛星放送のニュース＞

・１文が比較的短い．（１１．０語／文）
・無生物主語の文が多い
・口語的な表現がよく使われる．
・時制は「現在」の文が多い．

## ５．おわりに

ＮＨＫの衛星放送で試用されている機械翻訳システムの現況について報告した．また，実際に機械翻訳に供された英語ニュース文の分析結果も示した．衛星放送のニュースで使用することによって生ずる点を中心にシステムの問題点について触れた．さらに，現在構築中の英語ニュースデータベースおよびシステム改善の方向性について述べた．

英語ニュースデータベースはやっと利用可能となったばかりであり，本格的な利用はこれからである．これを用いて用語，慣用句，文型の抽出を行い，辞書の充実と文法の改良を進めていく予定である．

翻訳ソフトウェアはカテナ・リソース研究所のＳＴＡＲをベースに構築した．データ分析に当たって，同研究所の松田主任研究員に多大な協力を頂いたことを感謝します．

おわりに，研究の推進ならびに実用化に関して，ご指導を頂いた放送技術研究所の杉本所長，石田前次長，画像研究部の二宮部長，藤原前副部長に感謝します．

＜参考文献＞

１）相沢ほか「衛星放送ワールドニュースの英日機械翻訳」情報処理学会第４０回全国大会2F-1(1990)
２）中瀬「英日機械翻訳システムにおける解析手法について」情報処理学会自然言語処理研究会，NL67-7(1988)
３）加藤ほか「英日機械翻訳における固有名詞処理」情報処理学会第４０回全国大会2F-2(1990)
４）田中ほか「類語国語辞典を介した意味マーカー付与」情報処理学会第４０回全国大会6F-7(1990)
５）相沢ほか「英日機械翻訳のためのニュース用語辞書の構築」電子情報通信学会言語理解とコミュニケーション研究会，NLC91-6(1990)
６）浦谷「高速な複数文字列照合アルゴリズム：ＦＡＳＴ」情処論文誌，Vol.30,No.9(1989)